Modeling Feature Sharing between Object Detection and Top-down Attention

Dirk Walther[1*], Thomas Serre[2], Tomaso Poggio[2], Christof Koch[1]

[1] Computation and Neural Systems. California Institute of Technology, Pasadena, CA, 91125, * walther@klab.caltech.edu

[2] Dept. of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, 02139

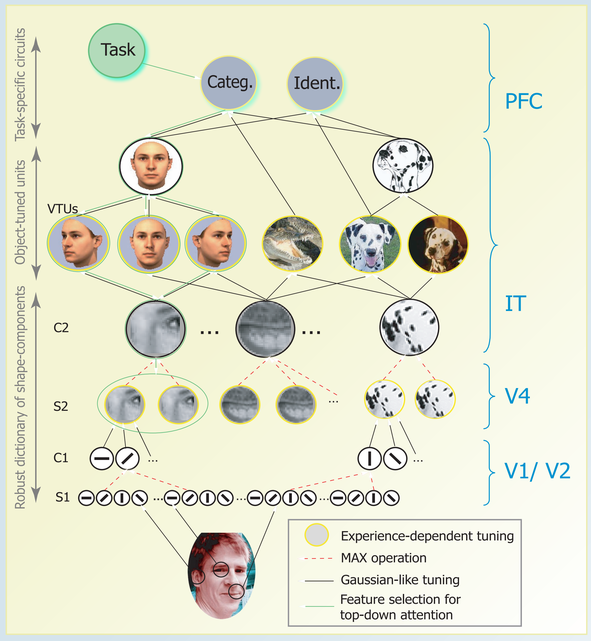
1046
VSS 05

Introduction

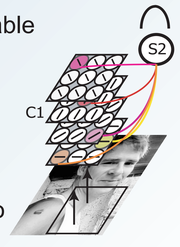
When performing visual tasks such as search for natural objects in a cluttered background, the attention system is biased from the top down for certain attributes of the targets. How is the task mapped to the particular features?

We propose that feedback connections in an object recognition system can serve this purpose. We demonstrate a computational implementation of such a system that, once trained for detecting faces, is capable of visual search for faces.

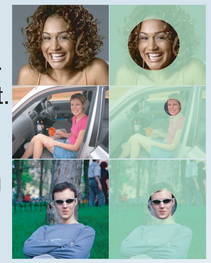
Model Architecture



- The model is based on the hierarchical feed-forward model of object recognition in cortex by Riesenhuber and Poggio [1] and its extension for feature learning at S2 by Serre and Poggio [2].
- A modified trace rule [3] selects a stable shape dictionary from snapshots of C1 activity (see also [4]).
- For top-down attention, feed-back connections from the abstract object representation down to the S2 level select a few S2 units. Their activity to a stimulus is used to bias spatial selection.



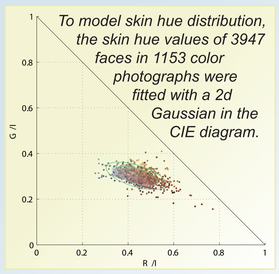
Data Sets



All training and test images are hand-labeled photographs from the internet. Two sets of 100 S2 features each were learned from 200 training images - set A from the entire training images; set B only from the face regions.

The recognition model was trained on the 200 training face images and 200 non-face training images. The ROC areas for in-dependent test sets were 0.989 for set A and 0.994 for set B.

Testing of top-down attention was done on a set of 179 images that contained between two and 20 faces, with a total of 593 faces.



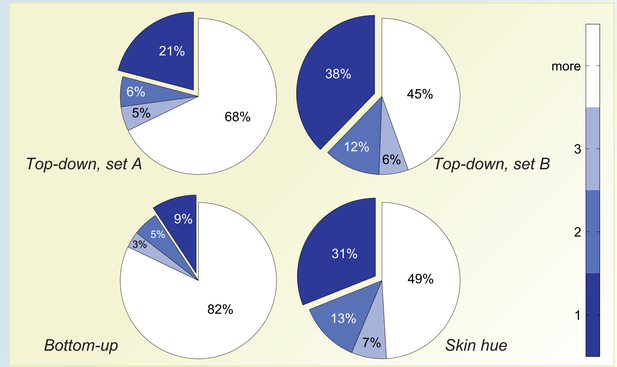
Top-down attention is compared to:

(1) The saliency-based model for bottom-up attention by Itti and Koch [5];

(2) top-down bias based on skin hue statistics.

Fixation Analysis

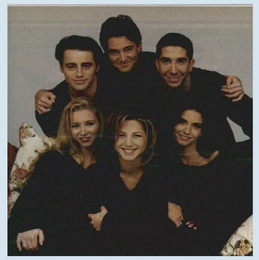
How many fixations (visits to the pixels of the activation maps in order of descending activity) does it take on average to find a face? For the n^{th} face in the image we measure the number of non-face fixations since finding the $(n-1)^{th}$ face in the image.

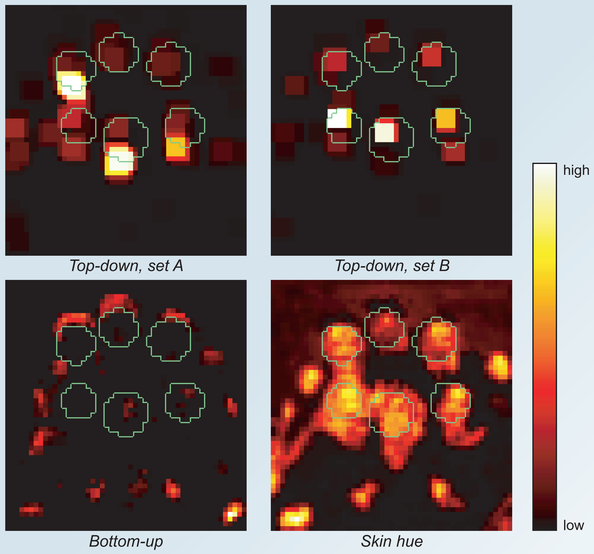


Region of Interest Analysis

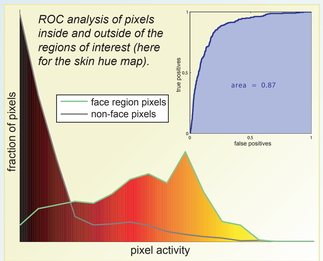
The activities of units in each map are separated into face and non-face pixels (based on ground truth).

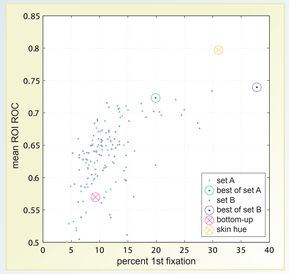
Separate activity value distributions are obtained for the two regions.





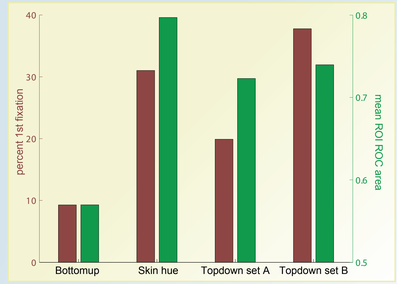
The area under the ROC curve for these activity value distributions, averaged over all test images, provides a performance measure for the top-down features, and bottom-up and skin hue maps, respectively.





The mean ROI ROC value correlates with the percent of faces that are found at the first fixation for the respective features. The best (highest percent first fixations) features are marked for both feature sets.

Summary of Results



According to both performance measures, top-down feature sets A and B perform better than bottom-up attention. Set A does not reach the performance of skin hue, while set B performs comparably, even outperforming skin hue in the number of faces that were attended in the first fixation. This is remarkable since top-down attention uses only grayscale versions of the images.

Conclusions

Features of intermediate complexity that are learned for the purpose of object recognition can be used effectively to guide top-down attention. Feedback connections in the visual hierarchy can provide a means of mapping an abstract task to a particular set of features that can be useful in solving the task.

In our computational implementation we have shown this behavior for faces. Both top-down feature sets performed better than mere bottom-up attention. Feature set A, which was derived from the entire training images, did not reach the performance of a skin hue detector. Set B, which was obtained from only the face regions of the training images, performed better than set A, reaching, and according to one performance measure, even surpassing the skin hue detector. The difference in performance between sets A and B suggests that some amount of guidance of the selection of training regions may be beneficial. In future experiments, we will assess the benefit of using bottom-up attention to guide feature learning.

Future work:

- Extend the implementation to several object categories;
- Implement a closed-loop system that verifies attended locations using the recognition sub-system;
- Make the system fully scale invariant.

References

1. Riesenhuber, M. and T. Poggio (1999), Nature Neuroscience, **2**(11): p. 1019-1025.
2. Serre, T. and T. Poggio (2005), VSS poster #744.
3. Foldiak, P. (1991), Neural Computation, 3: 194-200.
4. Sigala, R., T. Serre, T. Poggio, and M. Giese (2005), VSS poster #26.
5. Itti, L., C. Koch, and E. Niebur (1998), IEEE PAMI, **20**(11): p. 1254-1259.

Acknowledgements

Thanks to Xinpeng Huang for labeling the training and test images. The model figures are modified from Serre and Poggio, Cosyne 2005. This research is funded by grants from NSF, NIH, and NIMH.